A Verifiable Search Agent (ASA) for Political Elite Research

More at connorjerzak.com

Elite attribute labels are measurements.

Therefore they must be auditable, comparable, and time-defined.

LLMs show potential

But they break three dataset requirements:

Non-verifiable
No durable evidence trail (queries, snippets, URLs, timestamps).

Non-comparable
Open-world labels drift across countries and years (synonyms, translations, rebrands).

Temporal leakage
Coding year t using post-t sources contaminates "baseline" covariates.

(Ji et al., 2023)

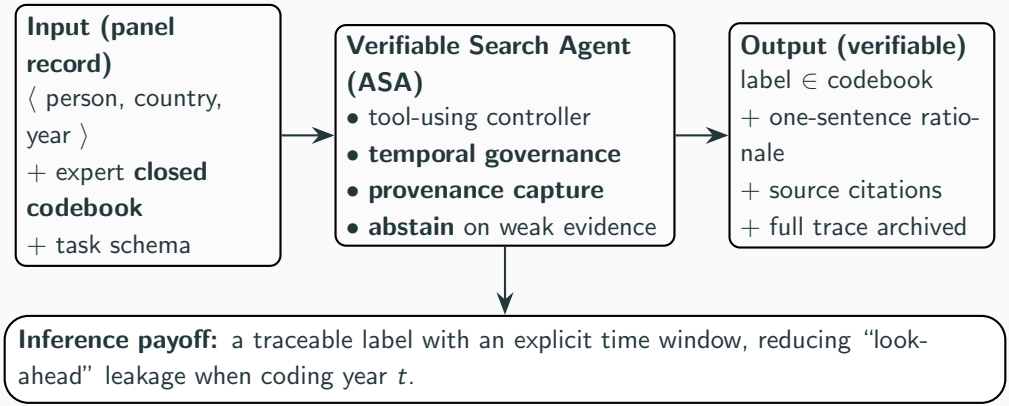
AI Search Agents Overcome Limitations

Verifiable	Comparable	Temporal leakage
Evidence-first outputs + explicit citations + full trace archive + low tool cost vs. APIs	Closed codebooks + conservative normalization + no new labels	"As-of" constraints + date checks + temporally constrained search

(Nakano et al., 2021; Yao et al., 2022)

Treat web retrieval as a
governed, auditable measurement instrument.

ASA in one picture

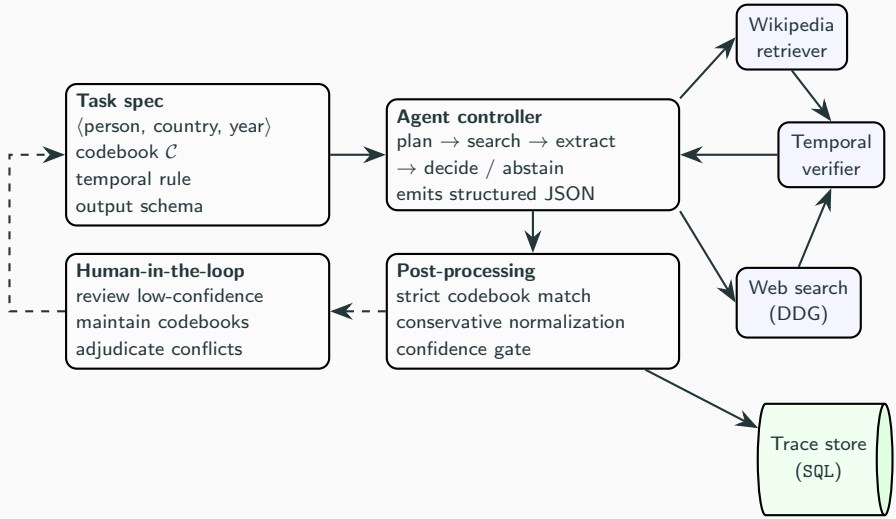


Commercial search tools vs. asa

Requirement	Commercial Grounded LLMs	ASA Protocol
Provenance	Opaque query generation; ephemeral links.	Full trace of all queries, raw snippets, URLs, tool calls.
Temporal Bounds	Biased toward the live/current web.	Strict "as-of" date constraints, checks.
Cost & Control	High search cost ($10 per 1k queries).	Modular, low-cost tool routing.

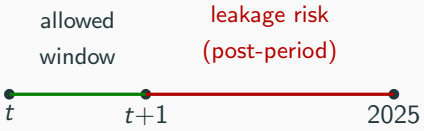
Therefore: ASA decouples the search execution from the reasoning step, lowering costs while guaranteeing an auditable evidence trail.

How ASA works



Temporal governance

- Target: attribute at year t (pre-period state).
- Common failure: use a source updated long after t, which summarizes later events.
- **Result:** your "baseline" covariate may encode post-t outcomes (party switches, coalitions).



(Kaufman, Rosset, and Perlich, 2012; Montgomery, Nyhan, and Torres, 2018)

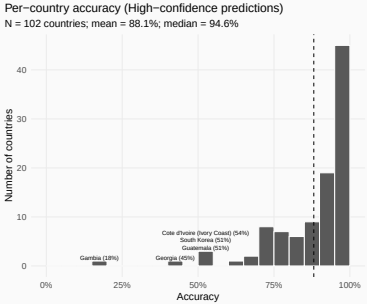
Temporal governance

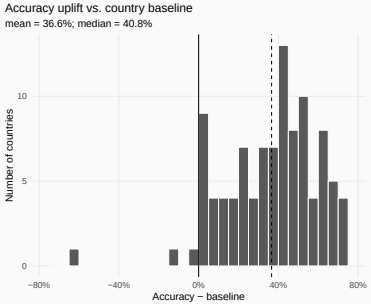
- Each run has an explicit "as-of" rule (e.g., sources published before $t+1$).
- When feasible: extract publication dates and enforce the cut strictly.
- When not feasible: warn and **abstain** rather than silently rely on post-period material.
- Full trace (including timestamps) is stored for later audit and alternative inclusion rules.

Case study: party affiliation labeling

- Task: label party affiliation for **34,618** leader–records across **114** countries.
- Closed world: **1,209** expert party labels (country-specific codebooks).
- Agent outputs: label $\in$ codebook + one-sentence rationale + citations + trace; otherwise abstain.
- Evaluation: score **high-confidence** predictions against expert codings.

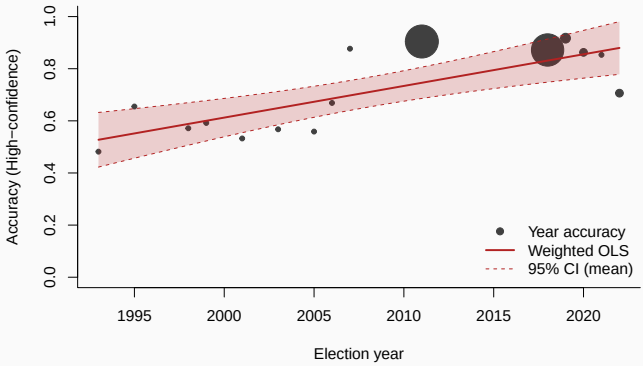
Results

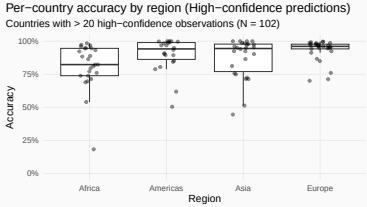




- **Accuracy (high-conf):** 0.860
- **Coverage gained:** +20.8% (12,898 new labels)
- **Withheld (low-conf):** 0.250 on average

Where performance varies





- Earlier years and sparse web records increase abstention and error risk.
- Name ambiguity and missing publication dates are recurring failure modes.

What “verifiable” looks like

Structured output (example)

```
{
  "label": "Party X",
  "confidence": "high",
  "rationale": "... (with citations)",
  "sources": [
    {"url": "...", "date": "..."},
    {"url": "...", "date": "..."}
  ]
}
```

Trace store enables

- audit any label quickly
- replicate and re-filter outputs
- route uncertain cases to humans

Takeaway: Verifiable agents turn LLM+web retrieval into replicable instruments for dataset construction.

References

Ji, Ziwei et al. (2023): **"Survey of Hallucination in Natural Language Generation".** In: *ACM Computing Surveys*, no. 12, vol. 55, pp. 1–38. DOI: 10.1145/3571730. URL: https://doi.org/10.1145/3571730.

Kaufman, Shachar, Saharon Rosset, and Claudia Perlich (2012): **"Leakage in data mining: Formulation, detection, and avoidance".** In: *ACM Transactions on Knowledge Discovery from Data*, no. 4, vol. 6, 15:1–15:21. DOI: 10.1145/2382577.2382579.

References ii

- Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres (2018): **"How conditioning on posttreatment variables can ruin your experiment and what to do about it".** In: *American Journal of Political Science*, no. 3, vol. 62, pp. 760–775.
- Nakano, Reiichiro et al. (2021): **WebGPT: Browser-assisted question-answering with human feedback.** arXiv: 2112.09332 [cs.CL]. URL: https://arxiv.org/abs/2112.09332.
- Yao, Shunyu et al. (2022): **ReAct: Synergizing Reasoning and Acting in Language Models.** arXiv: 2210.03629 [cs.CL].